Academy @

Stamicarbon

Intro to Data Science

▸ Day 1 → 5th Sep (13:00-17:00):
An Overview of Data Science Technologies

▸ Day 2 → 12th Sep (10:00-15:00):
Modern Applications of Data Science Methods

Trainer: Pedro V Hernández Serrano

Modern Applications of Data Science Methods

Pedro V Hernández Serrano
Workshop (Day 2) - Academy @Stamicarbon
12/09/2022

Stamicarbon

Lecture

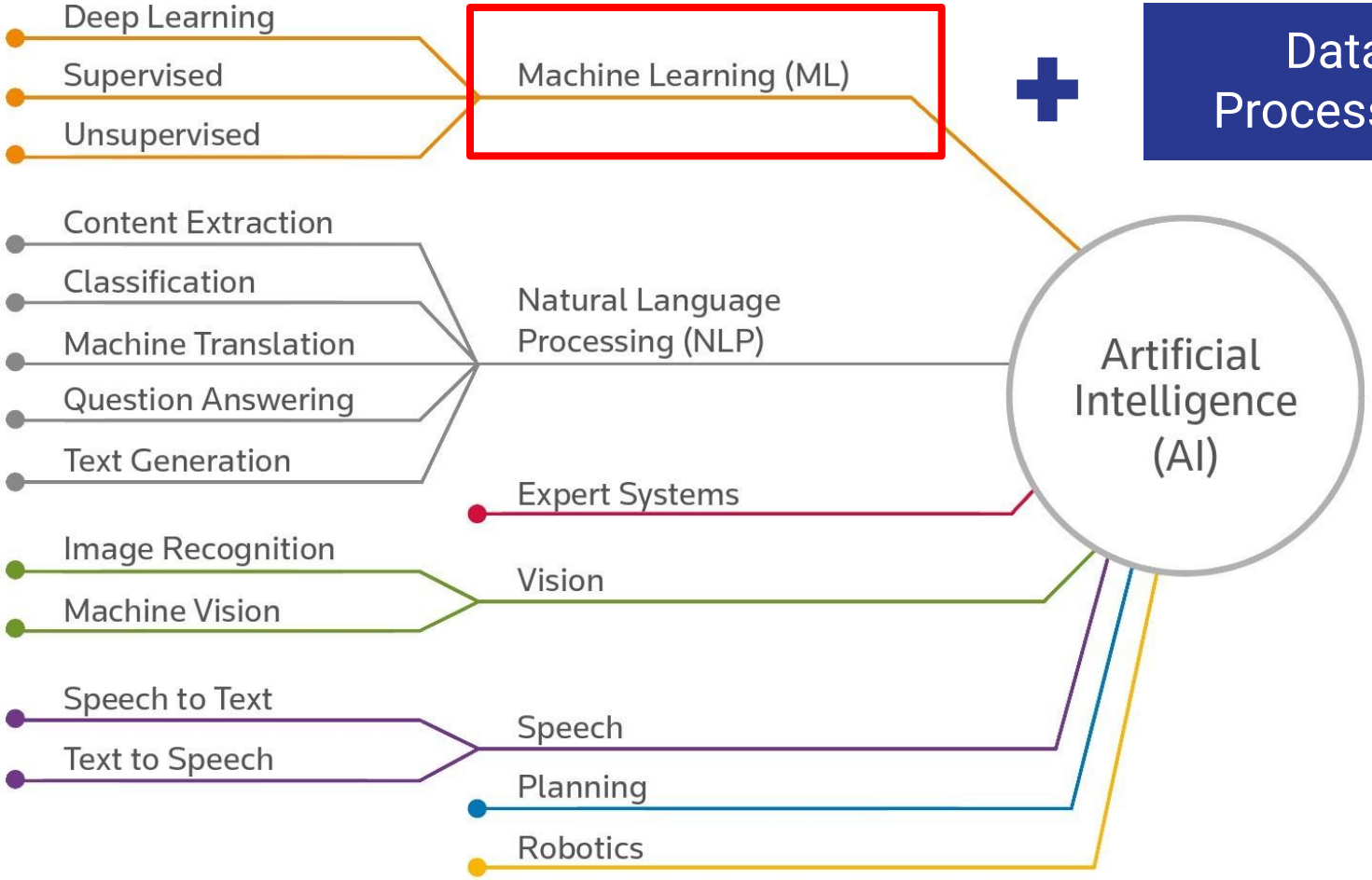
Day 2

- Data Science Methods
- Responsible Data Science
- Notebooks
  - Supervised
  - Unsupervised
  - Similarity

Goals

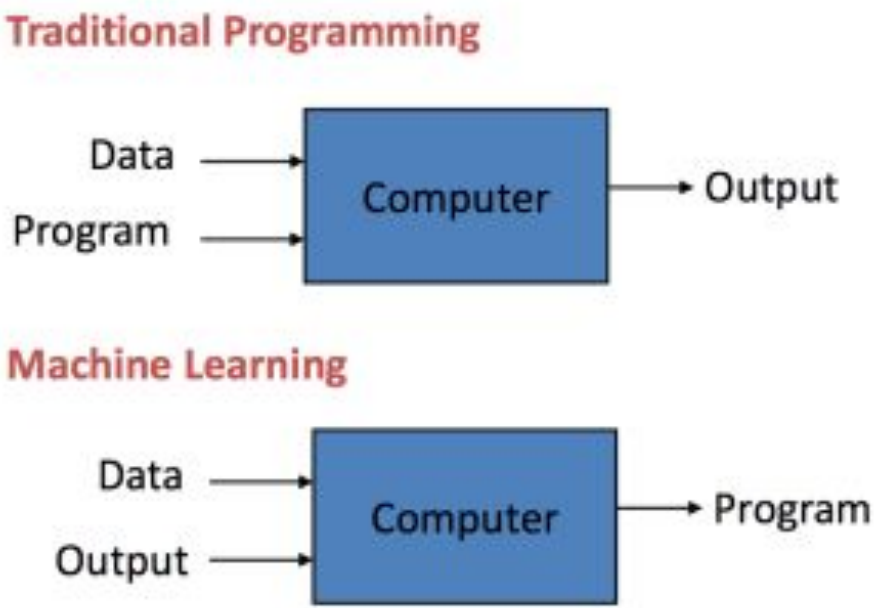
- Explore the taxonomy of methods in data science and AI
- Introduce the overall idea of what machine learning is.
- Get a sense of when deep learning can be helpful.
- Get an overview of the ethical, legal, moral and environmental considerations of machine learning.
- We will use open-source analytics tools
- We will learn through doing.

Data Science Methods

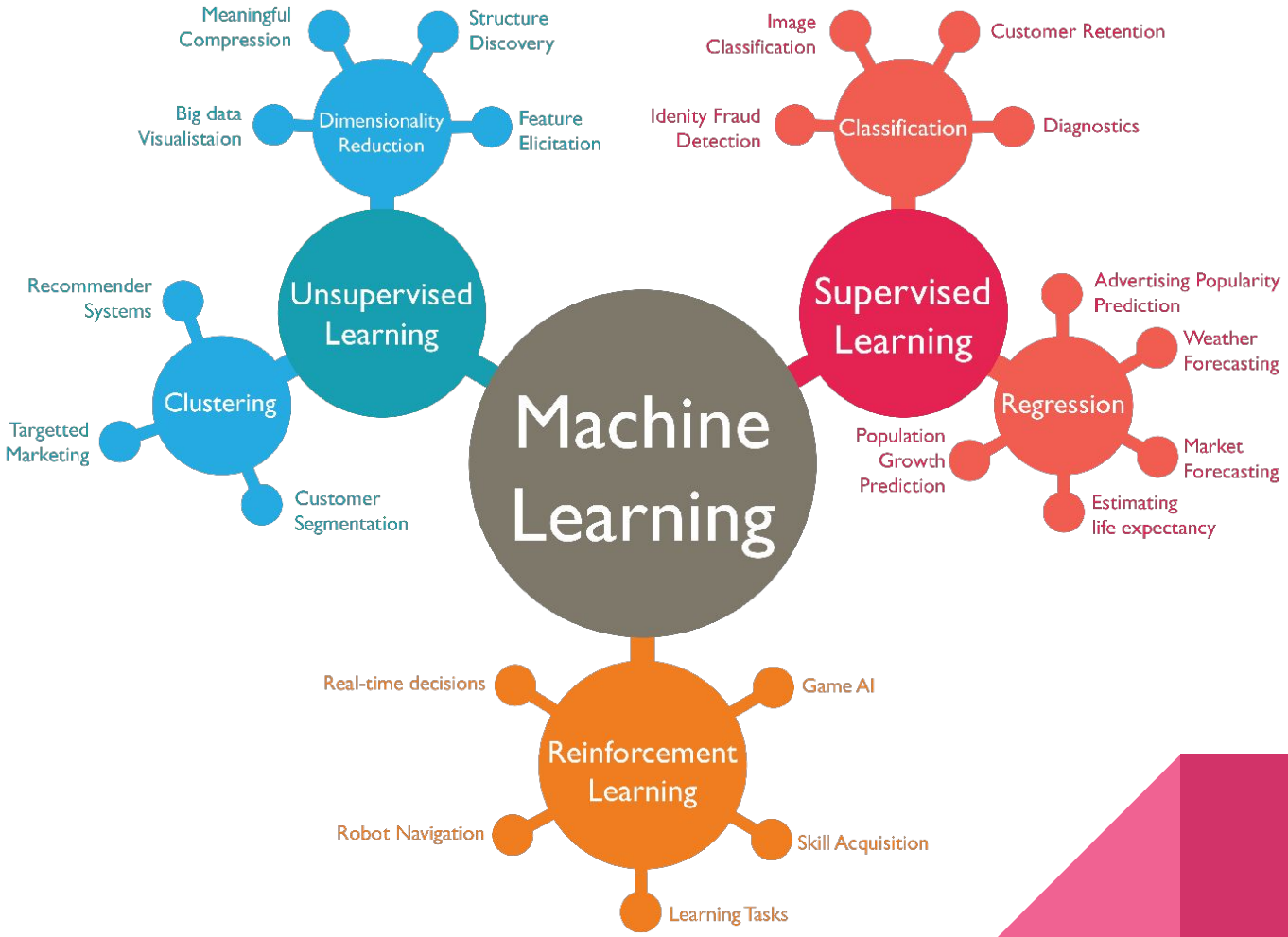


Data Processing

Machine Learning



Subfield of AI that focuses on the development of the computer programs which have access to data by providing the system with the ability to learn and improve automatically by finding patterns in the database without any human interventions or actions


Machine Learning
Unsupervised Learning
Dimensionality Reduction
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Clustering
Recommender Systems
Targetted Marketing
Customer Segmentation
Supervised Learning
Classification
Image Classification
Customer Retention
Idenity Fraud Detection
Diagnostics
Regression
Advertising Popularity Prediction
Weather Forecasting
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Reinforcement Learning
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

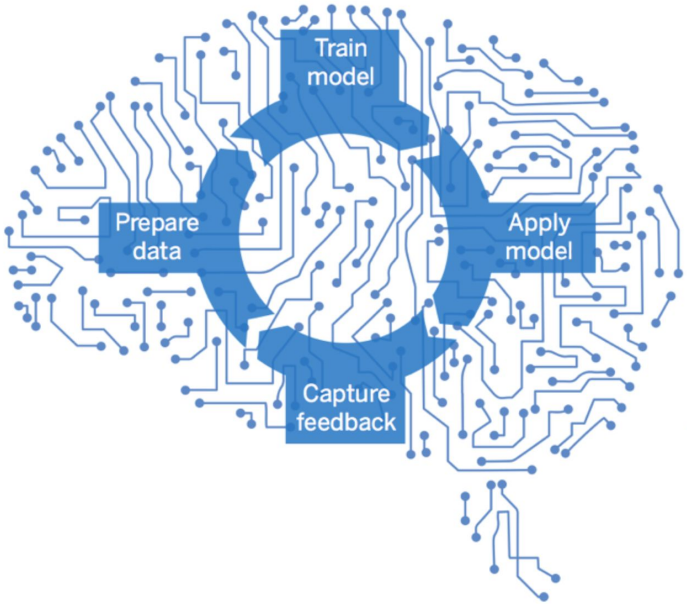
Deep Learning

Architectures

- Multi-layer perceptron (MLPs)
- Deep reinforcement learning (DRL)
- recurrent neural networks (RNN)
- Convolutional neural networks (CNNs)
- Generative adversarial networks (GANs)
- Transformers

Applications

- computer vision
- speech recognition
- natural language processing
- machine translation
- bioinformatics
- drug design
- medical image analysis
- climate science
- materials engineering
- board game programs

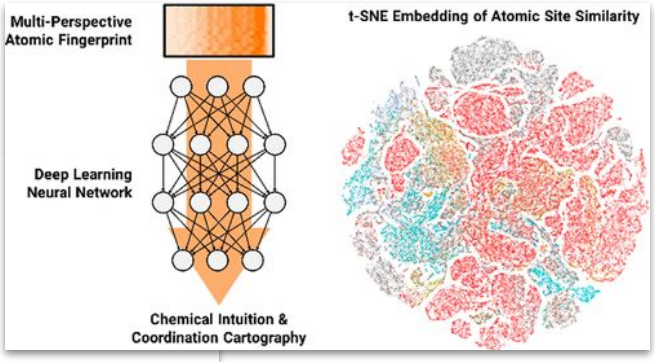


Sufficient high quality data labeled on a useful way

Swiss-army knife for any machine learning problem

The main difference is that we no longer follow the "kitchen-sink approach" with deep learning. Instead, the models do it for us.

Continuous Development



RETURN TO ISSUE | < PREV **ARTICLE** NEXT >

Crystal Structure Prediction via Deep Learning

Kevin Ryan*, Jeff Lengyel, and Michael Shatruk*

Cite this: *J. Am. Chem. Soc.* 2018, 140, 32, 10158–10168
Publication Date: June 6, 2018
https://doi-org.mu.idm.oclc.org/10.1021/jacs.8b03913

Article Views	Altmetric	Citations
13761	73	175

LEARN ABOUT THESE METRICS

- Paper: https://doi-org.mu.idm.oclc.org/10.1021/jacs.8b03913
- Docs: https://cctbx.github.io/
- Software: https://github.com/cctbx/cctbx_project

Resources

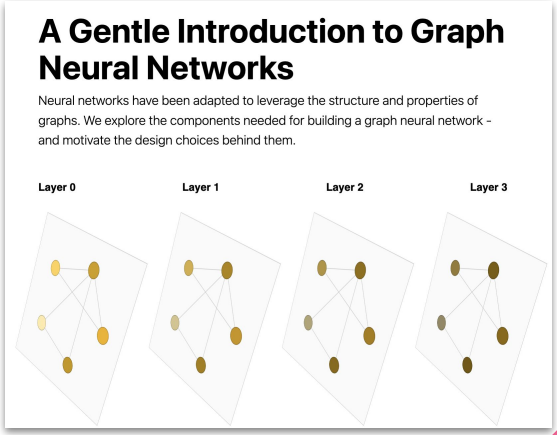
https://distill.pub/

Distill is dedicated to clear explanations of machine learning

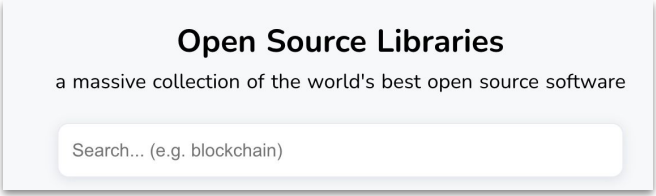
About Submit Prize Archive RSS GitHub Twitter ISSN 2476-0757

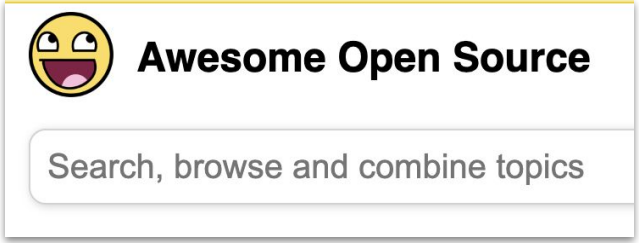
- **Graph Neural Networks**
https://distill.pub/2021/gnn-intro/

- **Generative Adversarial Networks**
https://distill.pub/2019/gan-open-problems/



Resources



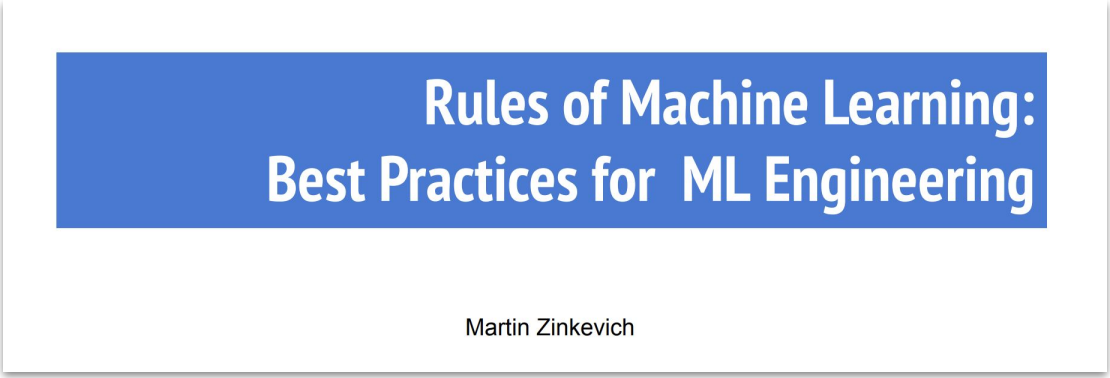


- OS Libraries: https://opensourcelibs.com/
- Awesome OS: https://awesomeopensource.com/
- ML Micorsoft: https://microsoft.github.io/ML-For-Beginners/#/
- IoT Microsoft: https://microsoft.github.io/IoT-For-Beginners/#/
- The ML Book: https://github.com/rasbt/python-machine-learning-book
- Detailed Notebook https://nbviewer.org/github/ageron/handson-ml/blob/master/index.ipynb
- To practice: https://github.com/sumantha-NTS/Fertilizer-Prediction

- **Repo:**
  github.com/pedrohserrano/data-science-technologies

Responsible Data Science

Best Practices



ML Rules:
https://martin.zinkevich.org/rules_of_ml/rules_of_ml.pdf

Azure ML pipelines:
https://docs.microsoft.com/en-us/azure/machine-learning/concept-ml-pipelines

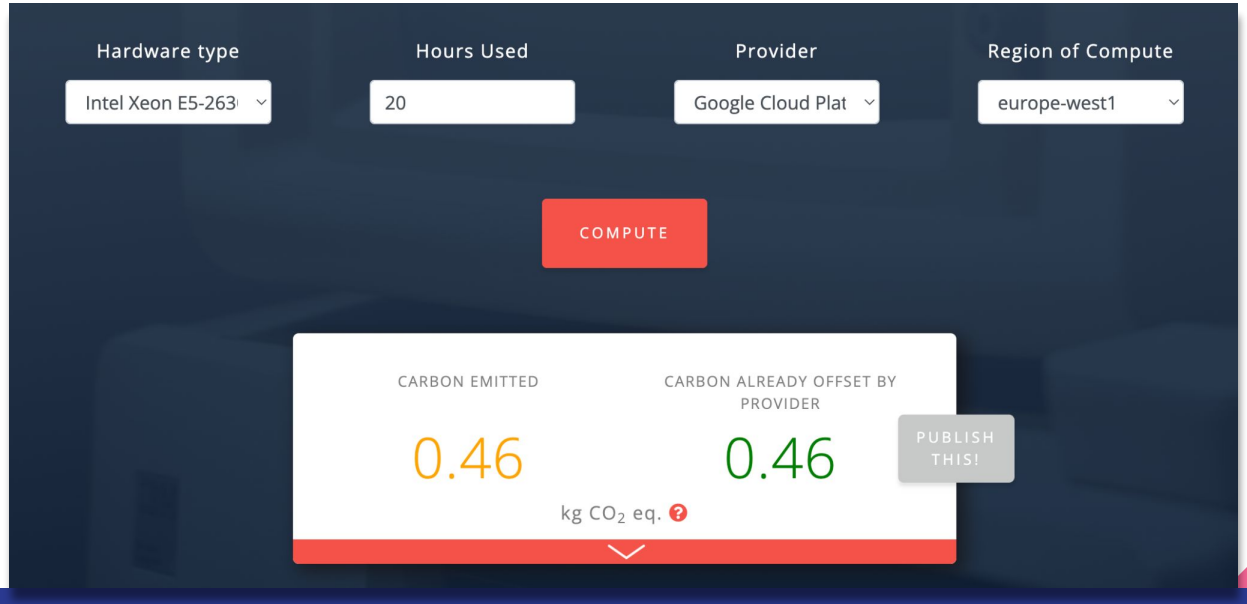
- Test the infra with the simplest model
- Always considering human-understandable features
- Development should be modular
- Documentation, documentation, documentation

Side effects

ML CO_2 IMPACT

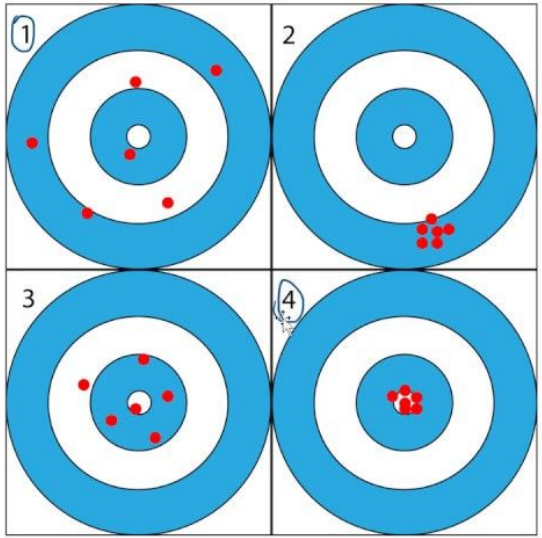
Machine Learning has a carbon footprint.

https://mlco2.github.io/impact/



Importance of Validity

Start with:
What do you want to measure?



1. Low Reliability, Low Validity
2. High Reliability, and Low Validity
3. Low Reliability, High Validity
4. High Reliability, High Validity

Source: USMLE Biostatistics: Precision, Accuracy, Internal & External Validity
https://www.youtube.com/watch?v=wD4S_CvsK18

Importance of Validity

Model predictions correlates with

Criterion 🎯	**Real data**	Is the accuracy evaluation appropriate? **AUC, Factor Analysis?**
External 📚	**Theory**	Do the outcomes correlate with the theory behind?
Content 🕵🏻	**Expertise**	No contamination on input, check with expert

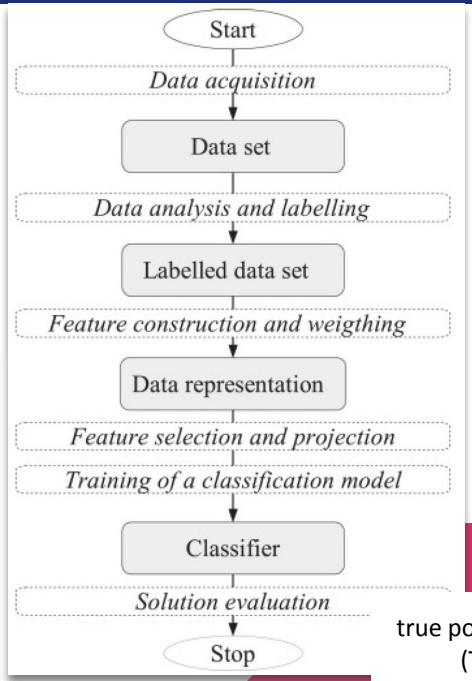
Am I measuring what I am supposed to be measuring?

Importance of Validity

Predicting hate speech in Twitter (an application of Sentiment Analysis) - Observational data

HaterNet - (2019) Pereira-Kohatsu et al.

- Used at Spanish National Office Against Hate Crimes
- Text classification + social network analysis
- 2 million tweets 6k tagged tweets



$\sum$

true positive rate (TPR)
false positive rate (FPR)

AUC 0.782

Accuracy of only Criterion Validity

Checking all 3 validities

ML Responsible Design

What happens if we only care about the criterion validity of the model?

FacePixelizer - (2020) Malimonov

PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models

NOT REAL PEOPLE

Human subject data

Ad Profiling

Input Data

Experience +
Customer Surveys +
Internet Customer Data

Scale: Millions of data points per user

Output Examples

- Personality traits
- Shopping activity
- Relationships (geographic + demographic + behavioural traits)

Want to use our data?

The Common Crawl corpus contains petabytes of data collected over 12 years of web crawling. The corpus contains raw web page data, metadata extracts and text extracts. Common Crawl data is stored on Amazon Web Services' Public Data Sets and on multiple academic cloud platforms across the world.

Access to the Common Crawl corpus hosted by Amazon is free. You may use Amazon's cloud platform to run analysis jobs directly against it or you can download parts or all of it.

You can search for pages in our corpus using the Common Crawl URL Index.

Get Started

Examples

Tutorials

Is it a blessing or a curse to have these methods as accessible, democratised, and easy to execute?